Course code	Data Science Programming	L	T	P	J	C
CSI3004		2	0	2	0	3
Pre-requisite		Syllabus version v.1.0				

Course Objectives:

1. To provide necessary knowledge on data manipulation and to perform analysis on the practical problems using statistical and machine learning approach
2. To generate report and visualize the results in graphical form using programming tool

Expected Course Outcome:

1. Ability to gain basic knowledge on data science
2. Gain the insights from the data through statistical inferences
3. Develop suitable models using machine learning techniques and to analyze its performance
4. Analyze on the performance of the model and the quality of the results
5. R tool for data Analysis and visualize the results
6. Demonstrate problem solving skills and provide solutions to real world problems

Module:1	Introduction	3 hours

Data Science: Basics – Digital Universe – Sources of Data – Information Commons – Data Science Project Life Cycle: OSEMN Framework

Module:2	Probabilistic Theory	4 hours

Probability Theory – Introduction – Conditional Probability – Bayes Rule – Gaussian Distribution – Inference of Gaussian

Module:3	Classification and Clustering	5 hours

Introduction to machine learning: Supervised, Unsupervised Learning – Regression: Linear Regression and Logistic Regression -- Classification Methods: K Nearest Neighbors, Naïve Bayes, Decision Trees - Clustering: k means, Hierarchical clustering

Module:4	Handling Data Using R	4 hours

R Objects, variables, datatypes, matrices, list, Control Structures, Functions, Data Frames, Reading and Writing Data File, Model Building		
Module:5	**Data Visualization in R**	**4 hours**
ggplot-univariate, bivariate, multivariate graph – time dependent graph – statistical models – histogram – box plot – heat map - scatter plot – legends – labeling		
Module:6	**Performance Evaluation**	**4 hours**
Model Evaluation Techniques: Hold out, cross validation - Prediction Errors: Type I, Type II - Loss Function and Error: Mean Squared Error, Root Mean Squared Error – Model Selection and Evaluation criteria: Accuracy, F1 score – Sensitivity – Specificity – AUC		
Module:7	**Data Analysis Using R – Case Study**	**4 hours**
Electricity consumption Data Analysis – Analysis of changes in pollution levels – Patient survival Analysis		
Module:8	**Recent Trends**	**2 hours**
	Total Lecture hours:	**30 hours**

Text Book(s)

1.	HadleyWickhmen, Garrette Grolemund, R for Data Science: Import, Tidy, Transform, Visualize and Model Data, OReilly, 2017
2.	Carl Shan, Henry Wang, William Chen, Max Song. The Data Science Handbook: Advice and Insight from 25 Amazing Data Scientists. The Data Science Bookshelf. 2016.

Reference Books

1.	Han, J., Kamber, M., Pei, J. Data mining concepts and techniques. Morgan Kaufmann. 2011
2.	Sergios Theodoridis, Konstantinos D Koutroumbas, Pattern Recognition, 4th Edition, Academic Press, Inc, 2009.
3.	James, G., Witten, D., T., Tibshirani, R. An Introduction to statistical learning with

	applications in R. Springer. 2013	
Mode of Evaluation: CAT / Assignment / Quiz / FAT / Project / Seminar		
List of Experiments		
1.	House rent prediction using linear regression	3 hours
2.	Medical diagnosis for disease spread pattern	3 hours
3.	Automate email classification and response	2 hours
4.	Customer segmentation in business model based on their demographic, psychographic and behavior data	3 hours
5.	Analysis of tweet and retweet data to identify the spread of fake news	2 hours
6.	Analyze crime data using suitable technique on reported incidents of crime based on time and location	2 hours
7.	Construct a recommendation system based on the customer transaction using Association rule mining	2 hours
8.	Perform analysis on power consumption data to suggest for minimizing the usage	2 hours
9.	Behavioral analysis of customers for any online purchase model	3 hours
10	Agricultural data analysis for yield prediction and crop selection on Indian terrain data set	3 hours
11.	Develop a recommender system for any real-world problem (when a user queries to find the university that offers Python, the system should display rank wise list of the university based on the review given by the customers)	3 hours
12.	Develop a business model to predict the trend in Investment and Funding	2 hours
	Total Laboratory Hours	**30 hours**
Mode of Evaluation: Project/Activity		

Recommended by Board of Studies	11-02-2021		
Approved by Academic Council	No. 61	Date	18-02-2021